DARIO Web Server

A free web server for the analysis of short RNAs from high throughput sequencing data

SEQUENCING EXPERIMENT AND ITS RESULTS

Illumina Genome Analyzer IIx
(http://www.illumina.com)

```
@HWI-EAS385:1:1:147:1370#0/1
TAGTTTATAAGACTGATGTTNATGGTNTGACGTCNTCTGCT
+HWI-EAS385:1:1:147:1370#0/1
QNT\^ab[T[TT`b\\b]Y[D][D]]D][F^ab]D^ba^a]
@HWI-EAS385:1:1:147:462#0/1
TAGCTTATAAGACTGATGATNATCGTNTGACGTCNTCTGCT
+HWI-EAS385:1:1:147:462#0/1
RT^bWbb]N]bb\a^^Y[D^DW[K]YD[YN^Zb]D^ba[_]
...
```

example output in fastq format

Sequence small RNAs **using a sequencing machine of your choice**.

The result will be a list of sequenced RNA molecules (reads) in a sequencer specific output file (typically in fastq format).

MAP READS TO A REFERENCE GENOME

Freely choose one of your favorite mapping tools to map your data to a reference genome. For example:

- segemehl
- BWA
- Bowtie
- SOAP
- ...

```
@HD     VN:1.0
@SQ     SN:chrI  LN:15072421
@SQ     SN:chrII LN:15279323
@SQ     SN:chrIII        LN:13783681
@SQ     SN:chrIV LN:17493785
@SQ     SN:chrV  LN:20919568
@SQ     SN:chrX  LN:17718854
@PG     ID:segemehl      VN:0.9.4-$Rev: 162 $ ($Date: 2010-10-15 12:48:37 +0200 (Fri, 15 Oct 2010) $)
GPL9269_GSM427346_GSE17153_102271 0       chrI     15070506 255     31M1D1M  *       0       0       ATTCTTAGTTGGTTGAGCGAT     *       NM:i:4   MD:Z     XN:i:1
GPL9269_GSM427346_GSE17153_102271 0       chrI     15063309 255     31M1D1M  *       0       0       ATTCTTAGTTGGTTGAGCGAT     *       NM:i:4   MD:Z     XN:i:1
GPL9269_GSM427346_GSE17153_384015 0       chrIV    3233310  255     21M      *       0       0       TGAGATCGTTCAGTACGGCAA     *       NM:i:0   MD:Z:21  XN:i:1
GPL9269_GSM427346_GSE17153_384015 0       chrIV    3233310  255     21M      *       0       0       TGAGATCGTTCAGTACGGCAA     *       NM:i:0   MD:Z:21  XN:i:1
GPL9269_GSM427346_GSE17153_384015 0       chrIV    3233310  255     21M      *       0       0       TGAGATCGTTCAGTACGGCAA     *       NM:i:0   MD:Z:21  XN:i:1
...
```

example output in sam format

PREPARE YOUR MAPPED READS FOR THE DARIO UPLOAD

Automatically convert the output to the needed bed format using **`map2bed.pl`.** Optionally you can create the bed file yourself.

`map2bed.pl` will not only create a bed formatted file, but also merge mapped reads to tags and zip the output file. This will **minimize the file size**, resulting in a **short upload time**.

(Note: It might be necessary to explicitly select 'SAM format' as output format when running your mapping tool.)

UPLOAD YOUR DATA TO THE DARIO WEBSERVER

Open the DARIO WebServer **http://dario.bioinf.uni-leipzig.de**

- Click on ***CLICK HERE TO START A NEW ANALYSIS JOB***
- Choose your reference species
- Choose upload file
- Optionally: choose a list of your own loci of interest
- Optionally: specify an e-mail address
- Click ***Proceed***

WAIT A MINUTE...

The job is automatically added to a queue and starts as soon as possible.
This page reloads every 30 sec and opens the result page when the job is done.

If you specified your e-mail address, you will get an e-mail with a link to the result page, as soon as the job is finished.

...DONE

After **~10 to 25 minutes** running time the result page is generated, containing the following sections:

- **Summary**
- **Quality Control**
- **Analysis**
- **Prediction**
- **User Annotation**
- **Download**

SUMMARY

Summary

Click here to bookmark this page for reference. This result will be available online at least 2 weeks from your jobs finish time.

You will find all downloadable files at the end of this page.

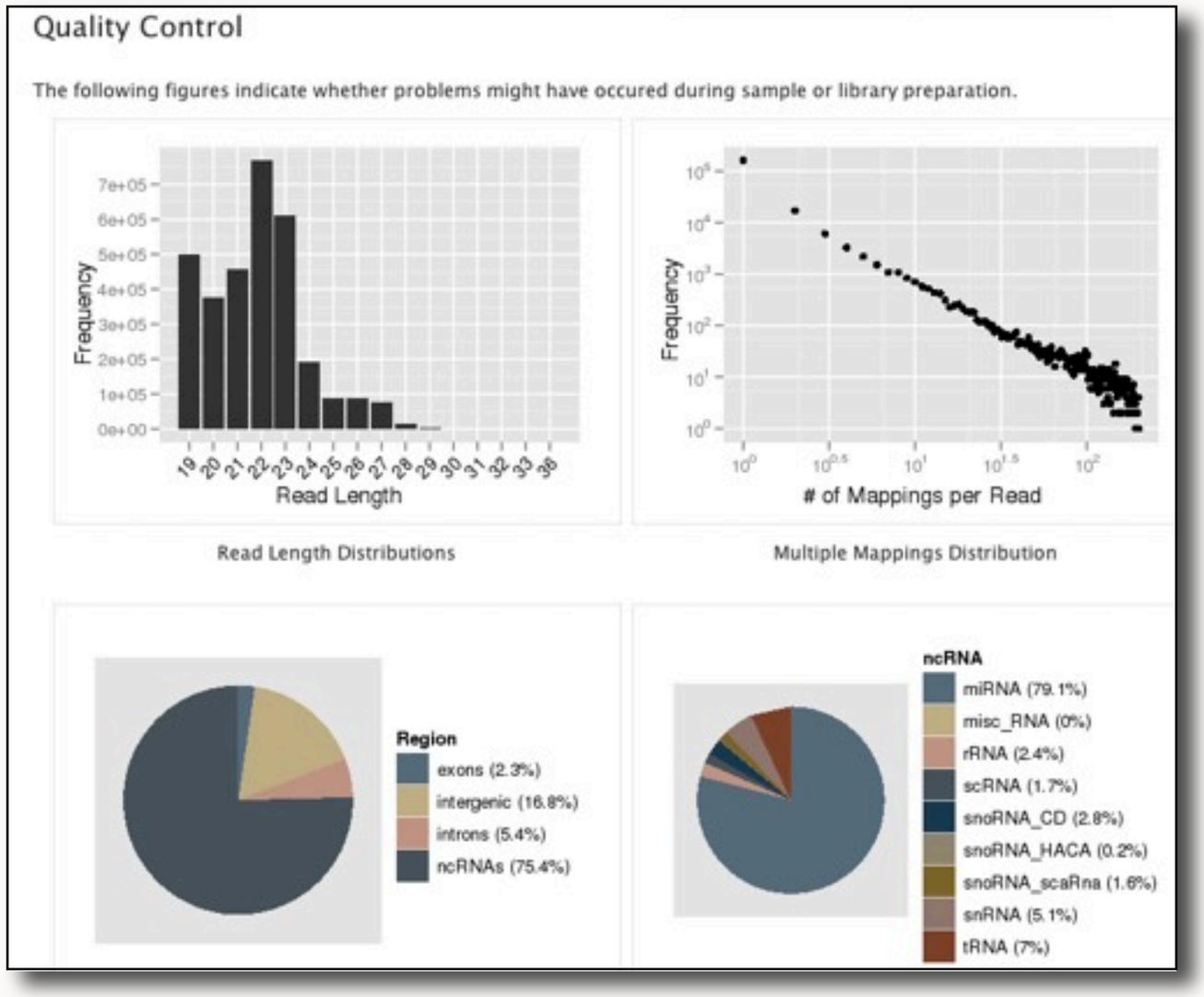
Job received at	2011-01-24 17:40:51
Job finished at	2011-01-24 17:50:04
Uploaded file name	GSM450600.bed.gz
# of uploaded mapping loci	655,276
Total # of reads	6,632,854
Total # of tags	351,359

The **Summary** contains some basic information about the job, e.g.:

- The date and time you uploaded your file
- The number of mapped loci you uploaded
- The number of reads and the number of tags*

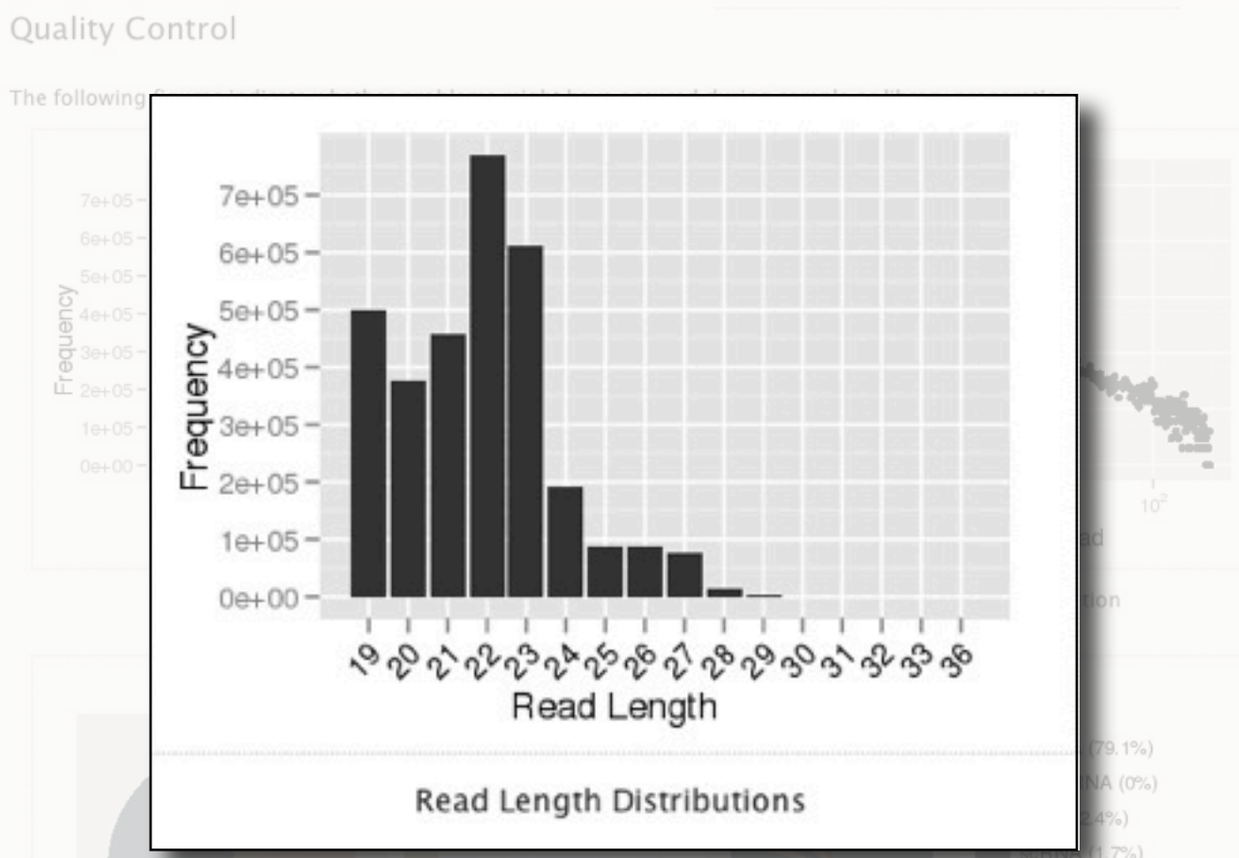
*A tag is defined as a RNA sequence that occurs at least once in a set of sequencer reads. Thus a tag typically corresponds to several identical reads.

QUALITY CONTROL



The **Quality Control** gives a first impression on how good your experiment performed.

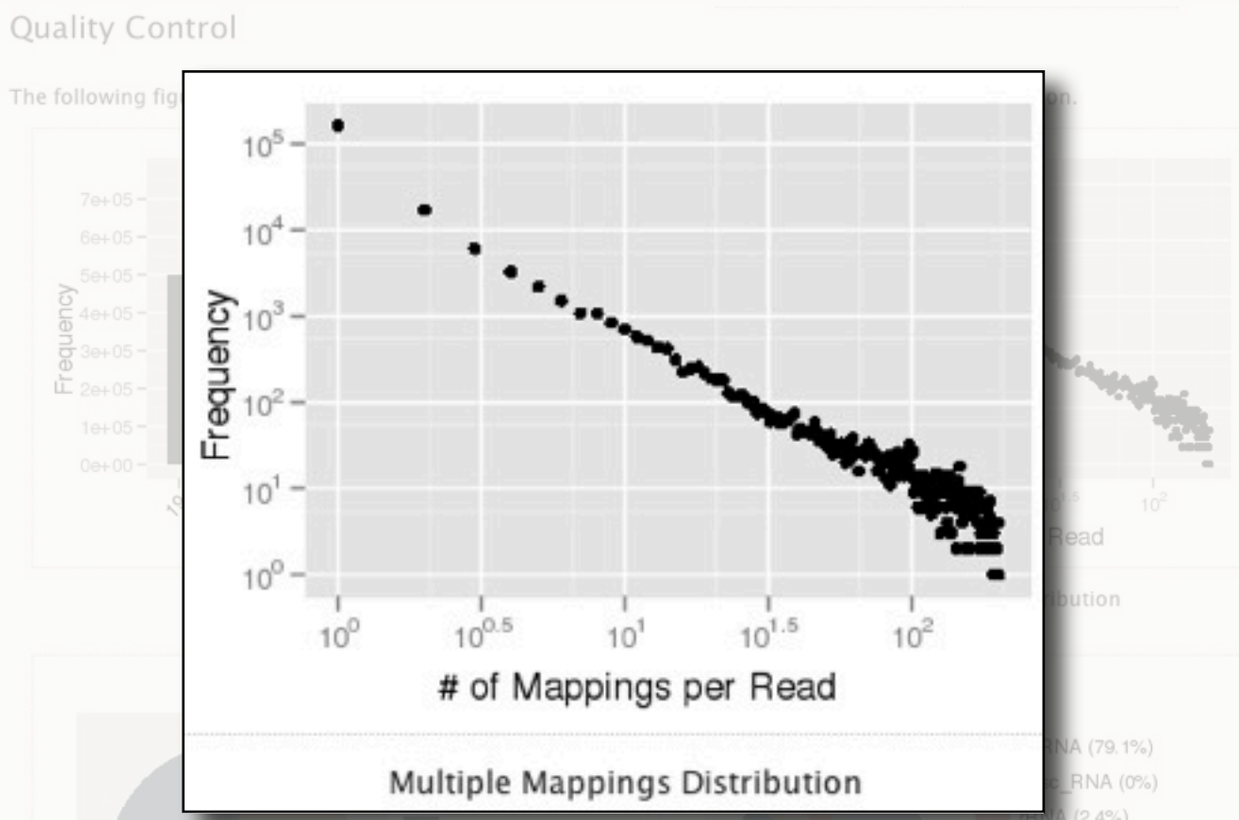
QUALITY CONTROL
READ LENGTH DISTRIBUTION



Read Length Distribution

Here, you can easily assess if the short read sequencing experiment performed properly. When using a short RNA preparation protocol, one would expect a peak at the length of 22.

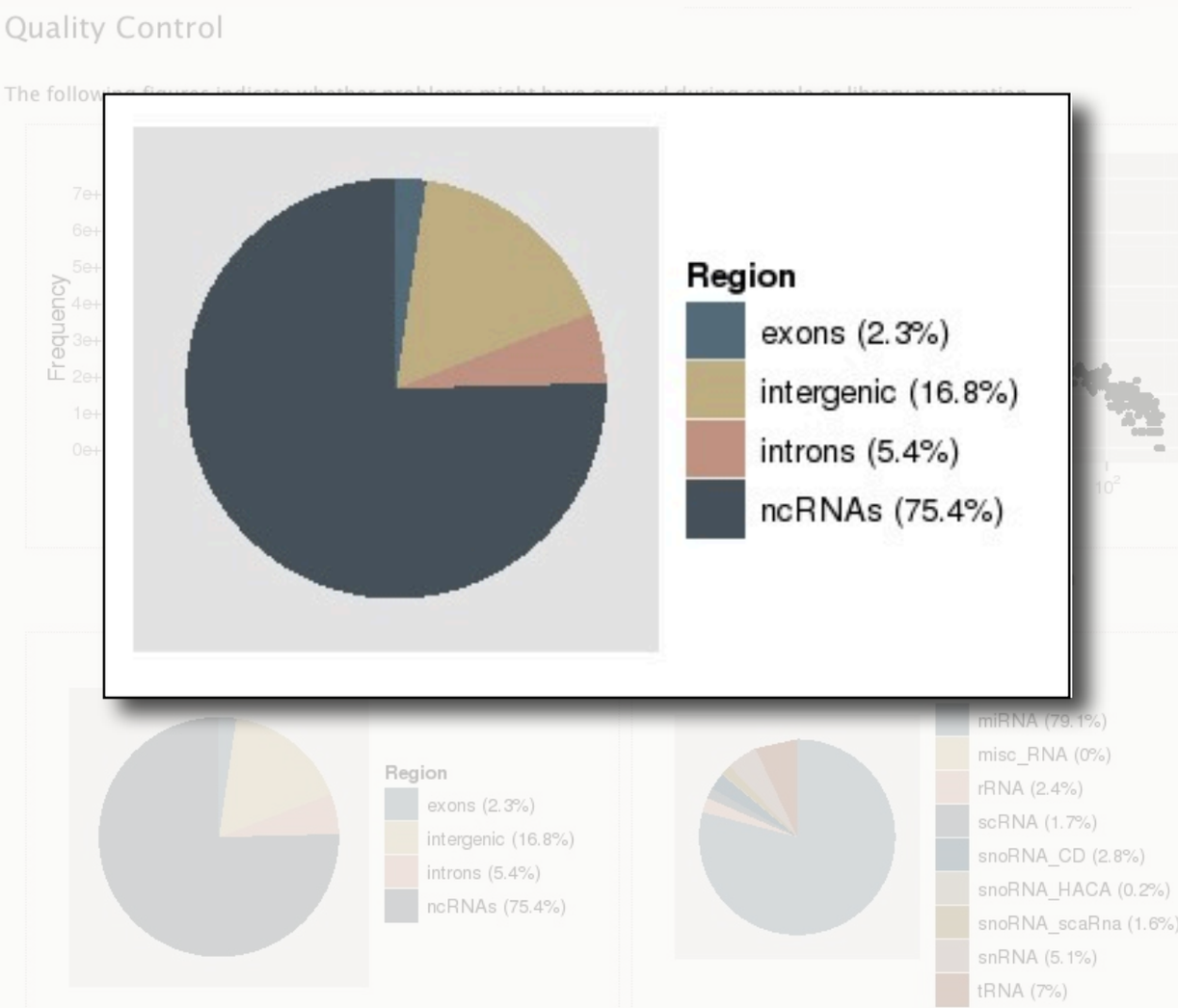
QUALITY CONTROL
MULTIPLE MAPPINGS DISTRIBUTION



Multiple Mappings Distribution

Most of the reads should map to a unique position within the reference genome. If this plot gives any doubt to this statement, you should have a deeper look on the quality values of your experiment.

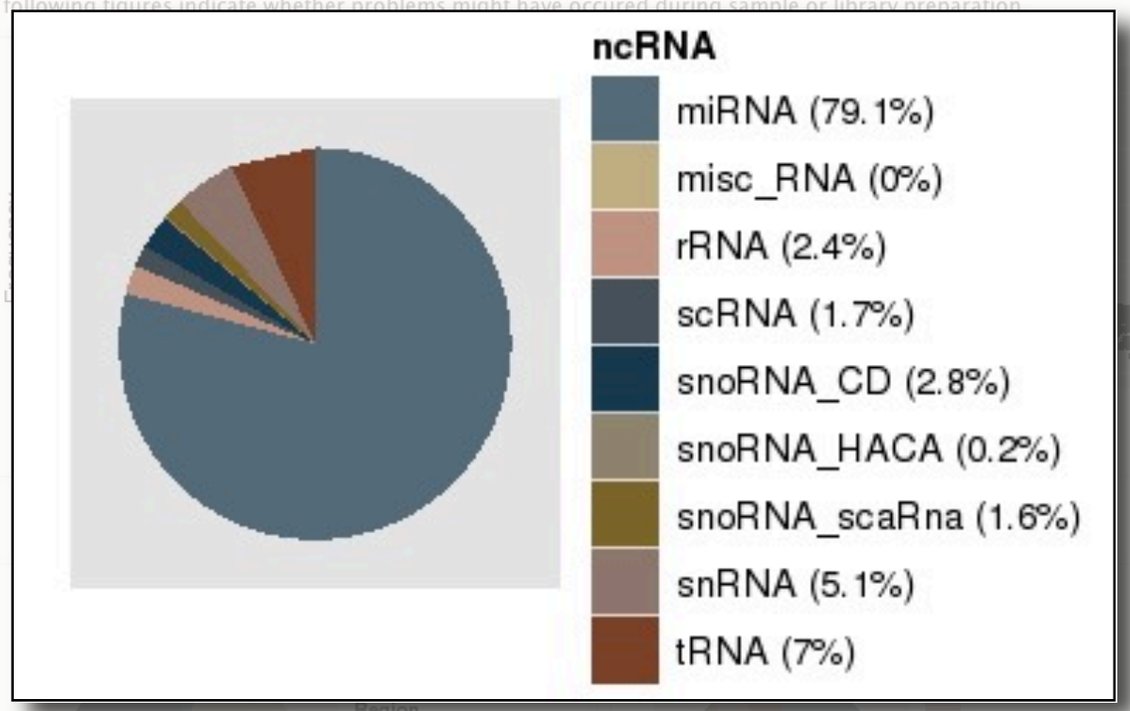
QUALITY CONTROL
MAPPING REGIONS WITHIN THE GENOME



Mapping Region

This plot shows the fraction of reads mapping to exons, introns, ncRNAs or not-annotated regions. The majority of the reads should map to annotated ncRNAs, if a short RNA protocol was used.

QUALITY CONTROL
READS MAPPING TO NON-CODING RNA LOCI



ncRNA Overlap

This picture gives an overview of the overall ncRNA expression in the experiment. Typically, miRNAs are the majority.

ANALYSIS

Analysis

This table summarizes the quantification of of the different ncRNA classes. Click on "View List" for detailed information for individual ncRNA species.

ncRNA Class	Reads	Reads (normalized)	# of Genes	Table
miRNA	1,141,865	816,175.5	661	View List
snoRNA_CD	50,991	29,031.09	181	View List
snoRNA_HACA	2,125	2,019.533	85	View List
tRNA	433,796	72,332.05	537	View List
scRNA	286,672	17,815.27	900	View List
snRNA	294,027	52,661.4	712	View List
rRNA	92,582	24,588.18	389	View List
snoRNA_scaRna	16,770	16,752.33	19	View List
misc_RNA	105	95.00559	8	View List

The tables in the **analysis section** are itemized by the types of ncRNAs.

ANALYSIS
EXAMPLE: MICRORNAS

ncRNA Class	Reads	Reads (normalized)	# of Genes	Table
miRNA	1,141,865	816,175.5	661	View List
snoRNA_CD				

Expression of ncRNA: miRNA

Go Back

miRNA expression sorted with location. Click the header to sort the table with respect to another column. Click twice to reverse sort order.

Chromosome	Start Loci	End Loci	Strand	ID	RPM	Reads	Reads (normalized)	Visualization
chr10	100144965	100145054	-	hsa-mir-1287	2.44e-01	30	30.00	View at UCSC
chr10	103351164	103351244	+	hsa-mir-3158-1	1.90e-01	42	21.00	View at UCSC
chr10	103351164	103351244	-	hsa-mir-3158-2	1.90e-01	42	21.00	View at UCSC
chr10	104186259	104186331	+	hsa-mir-146b	1.20e+01	1.203	1.202.00	View at UCSC

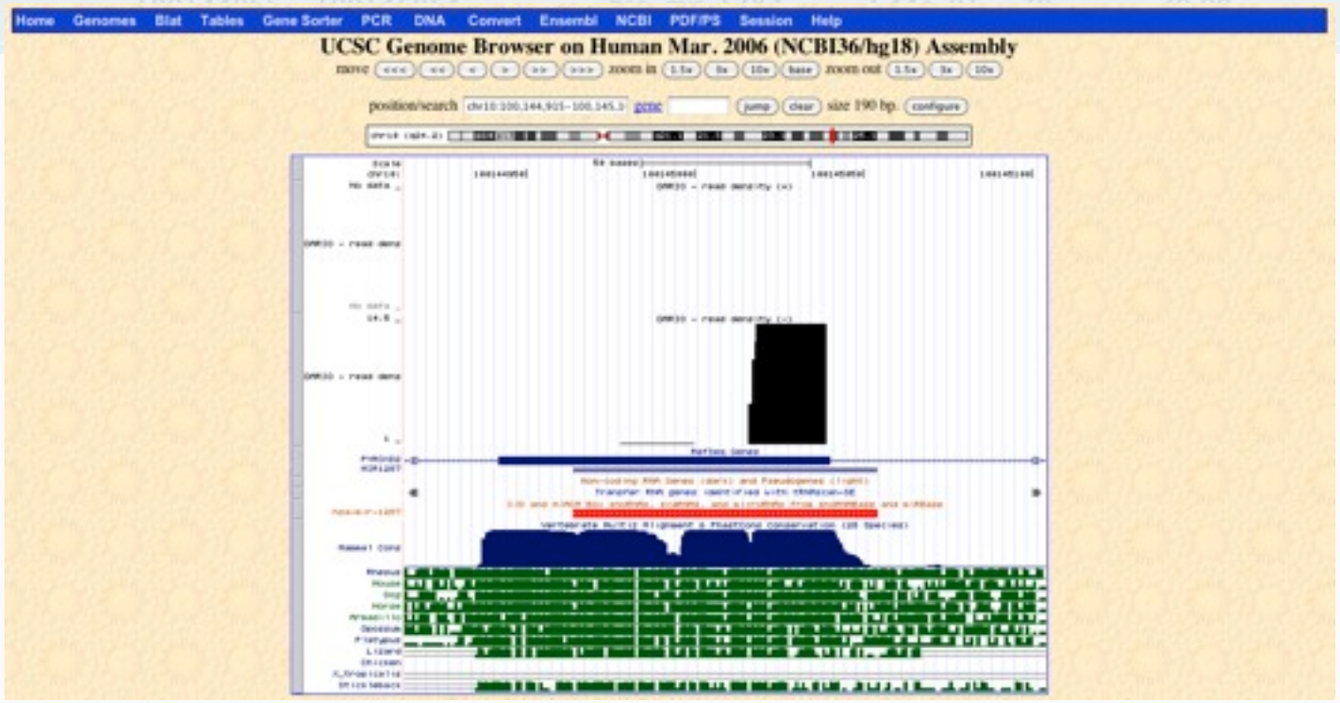
In the expression lists contains:

- the ncRNA loci (chromosomal position)
- the ncRNA ID
- the number of reads overlap with the loci
- the number of overlapping reads, normalized for multiple mappings
- the RPM (Reads Per Million) normalized expression for each expressed ncRNA

ANALYSIS
EXAMPLE: MICRORNAS

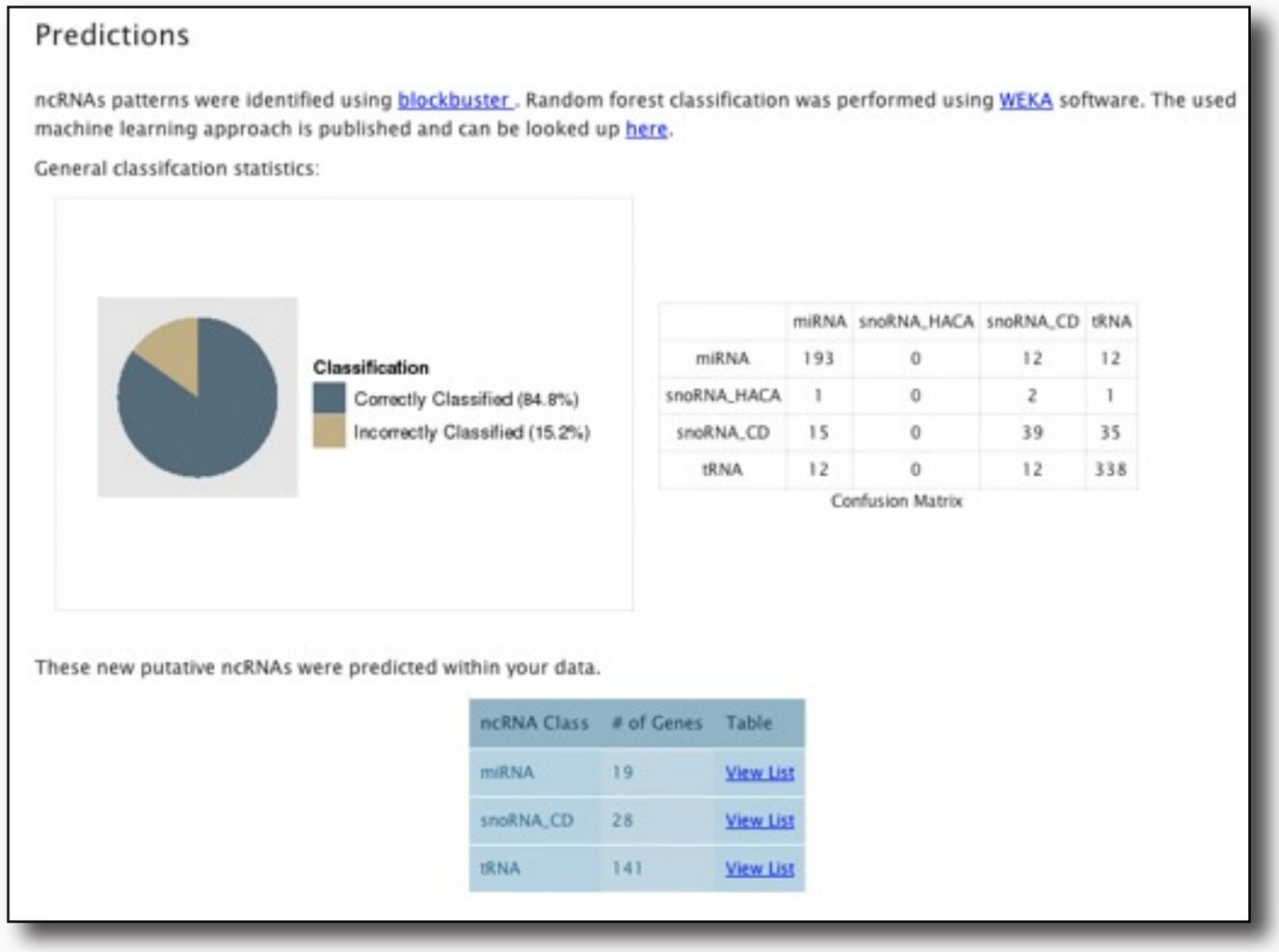
ncRNA Class	Reads	Reads (normalized)	# of Genes	Table
miRNA	1,141,865	816,175.5	661	View List
snoRNA_CD	50,991	29,031.09	181	View List

Chromosome	Start Loci	End Loci	Strand	ID	RPM	Reads	Reads (normalized)	Visualization
chr10	100144965	100145054	-	hsa-mir-1307	7.44e-01	19	19.00	View at UCSC



Click ***View at UCSC*** to be forwarded to the UCSC Genome Browser showing the region of the ncRNA. Take a look on the **expression pattern**.

PREDICTION



Predictions

ncRNAs patterns were identified using blockbuster . Random forest classification was performed using WEKA software. The used machine learning approach is published and can be looked up here.

General classifcation statistics:

	miRNA	snoRNA_HACA	snoRNA_CD	tRNA
miRNA	193	0	12	12
snoRNA_HACA	1	0	2	1
snoRNA_CD	15	0	39	35
tRNA	12	0	12	338

Confusion Matrix

These new putative ncRNAs were predicted within your data.

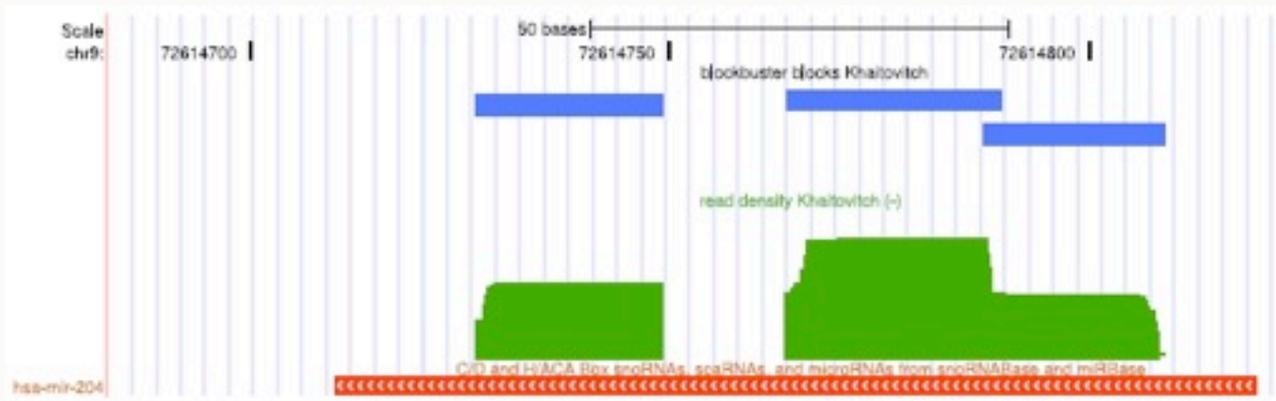
ncRNA Class	# of Genes	Table
miRNA	19	View List
snoRNA_CD	28	View List
tRNA	141	View List

The upper part gives information about the sensitivity of the predictions on the uploaded dataset. Below the predicted loci are shown.

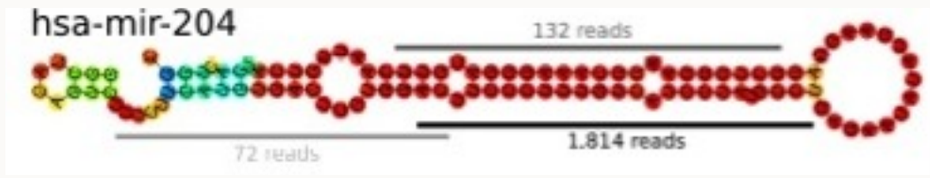
PREDICTION
THE IDEA

Short RNA reads map to ncRNAs forming specific patterns (Langenberger et al.).

This is best known for miRNA, where the reads form high stacks right above the annotated mature and the mature* loci.



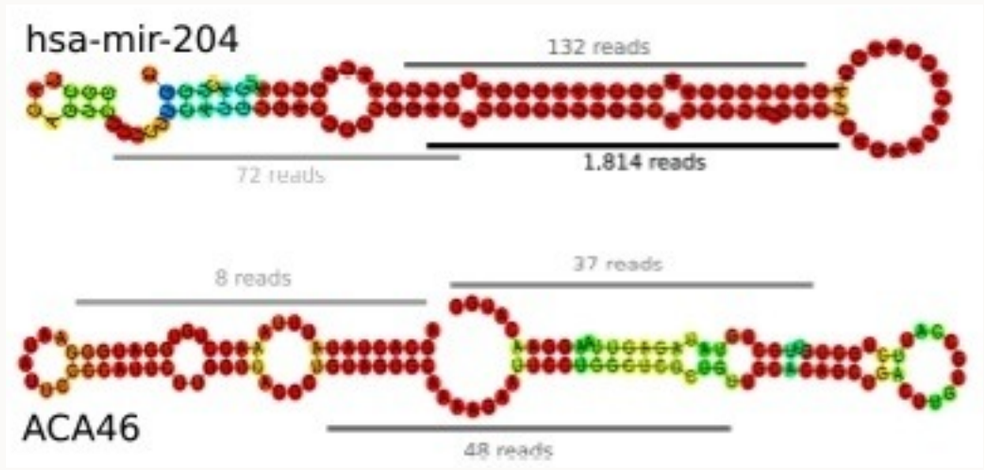
read pattern for hsa-mir-204 in the UCSC genome Browser

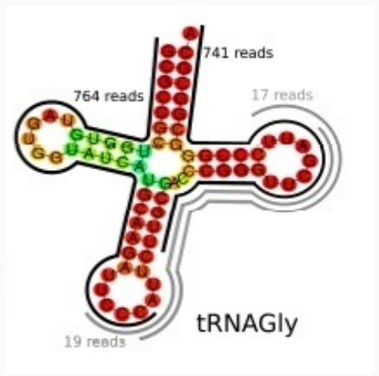


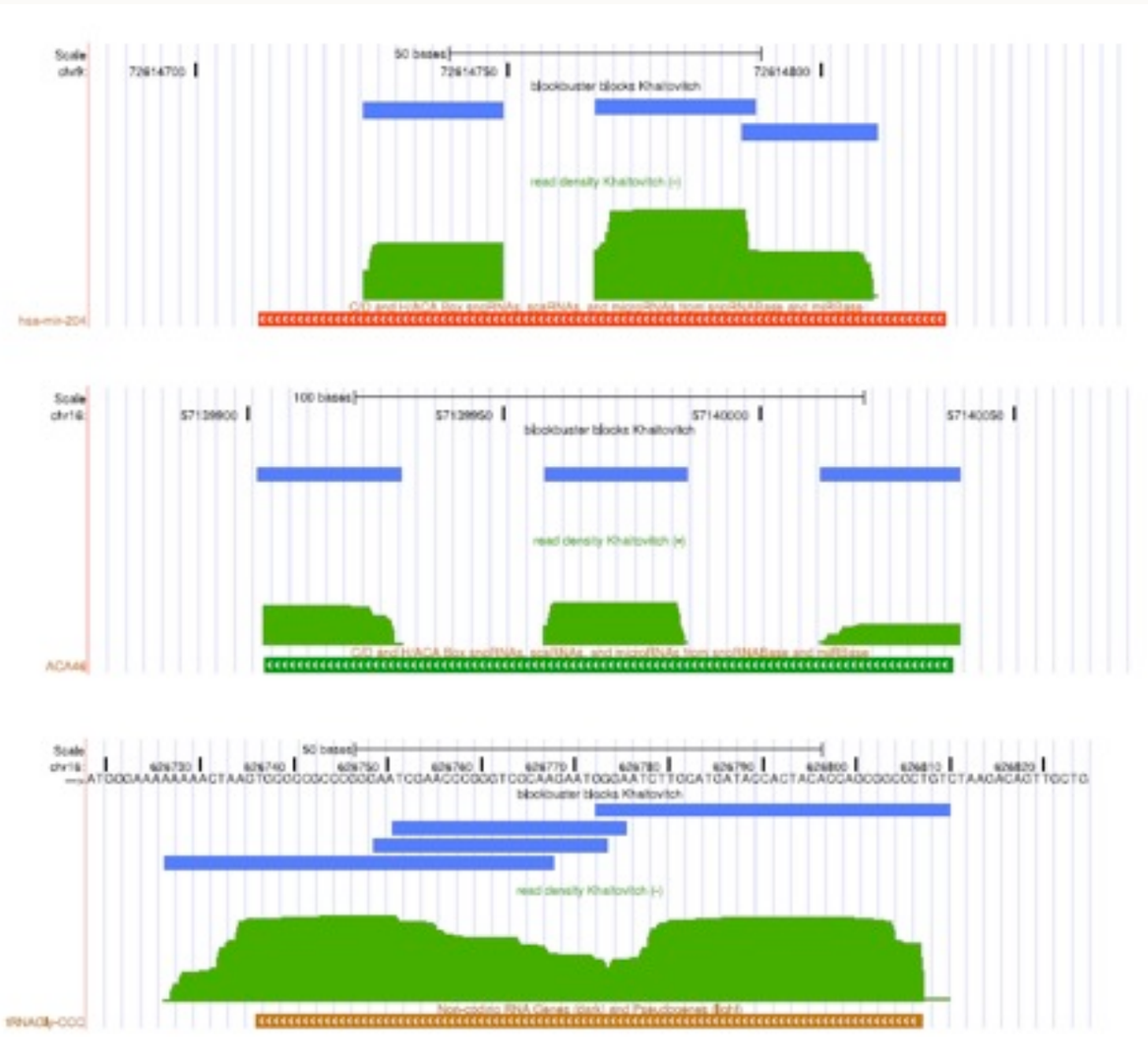
read pattern on the folded microRNA

PREDICTION
THE IDEA

Other classes of ncRNAs show different expression patterns:

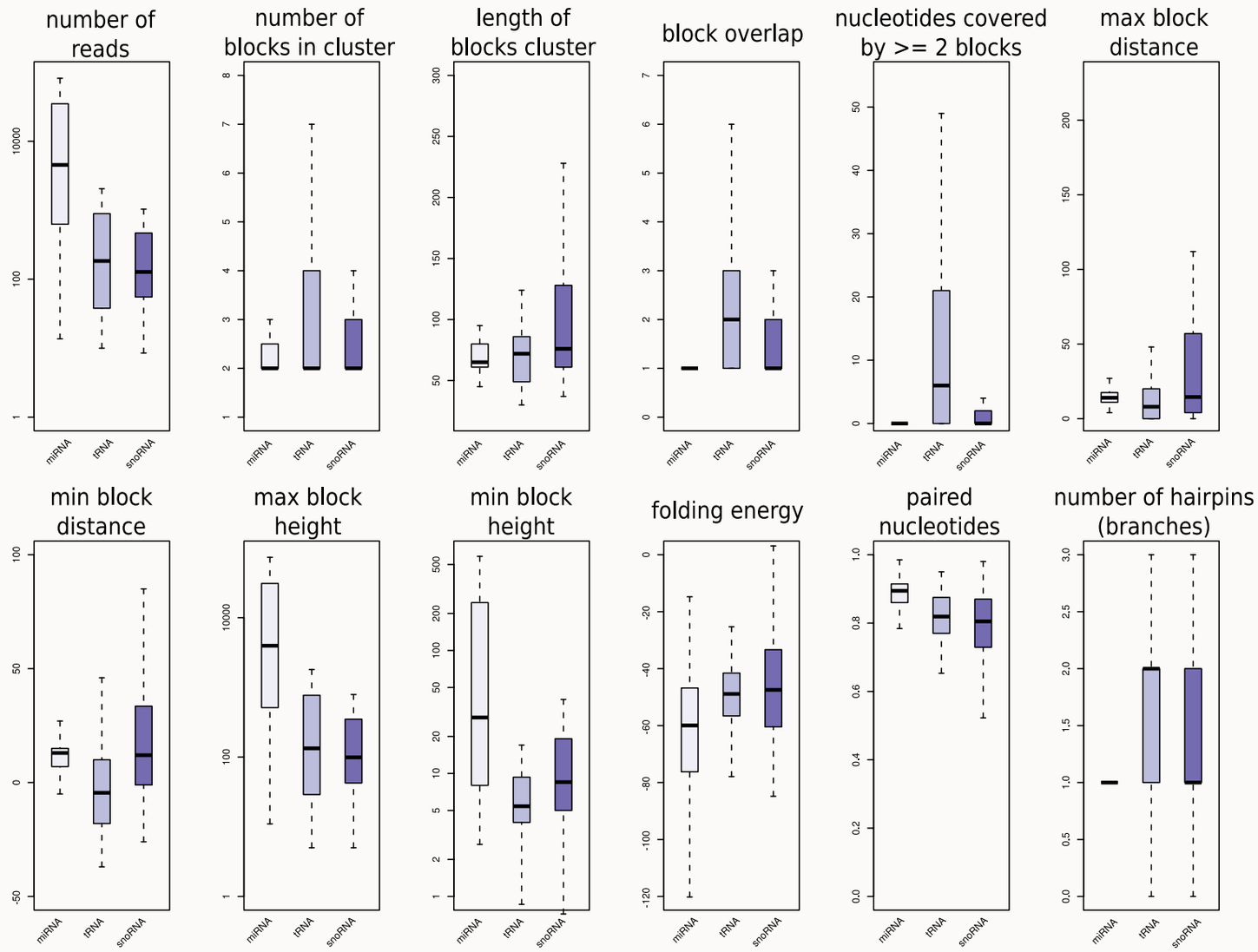






PREDICTION
THE IDEA

Different features are used to separate the classes, using a RandomForest classifier:



PREDICTION
THE OUTPUT

New ncRNAs candidates are identified in the uploaded data and ranked by their scores.

Predictions of ncRNA: miRNA

Go Back

Sorted miRNA predictions descending with score. Click the header to sort the table with respect to another column. Click twice to reverse sort order.

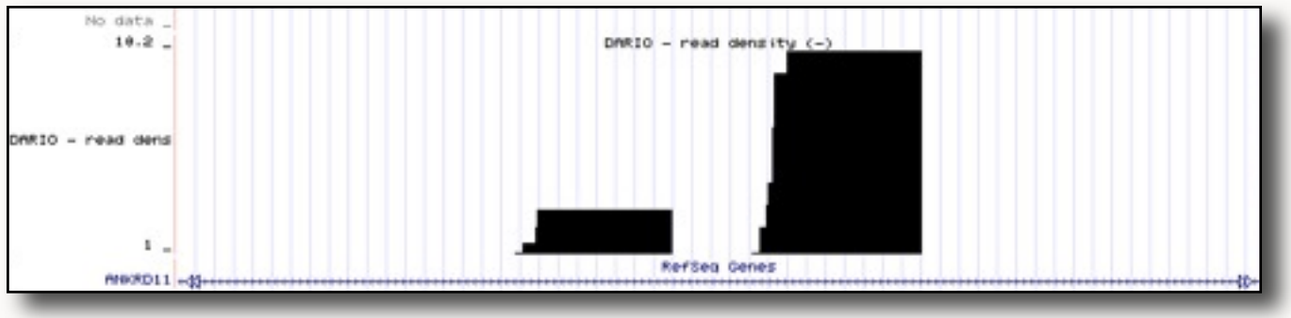
Chromosome	Start Loci	End Loci	Strand	ID	Score	RPM	Reads	Reads (normalized)	RNAz Validation	Visualization
chr11	67984196	67984253	-	miRNA_94	1	1.62e-01	14	14.00		UCSC
chr14	65007583	65007643	-	miRNA_1	1	1.27e+00	211	106.00		UCSC
chr11	62091059	62091114	-	miRNA_76	0.99	1.83e-01	14	14.00		UCSC

Futhermore, we overlap the predicted loci with RNAz screens. RNAz uses conservation and secondary structure properties to predict regions likely to form functional RNA structures.

PREDICTION
THE EXAMPLES

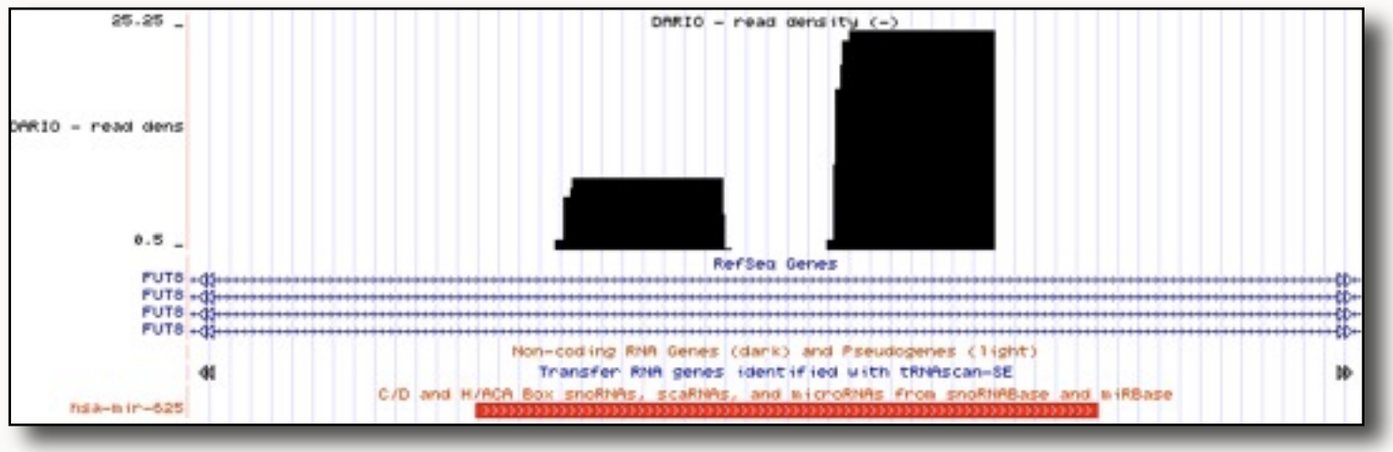
Currently unknown microRNA candidates predicted by DARIO:

new candidate within an annotated intron of a gene

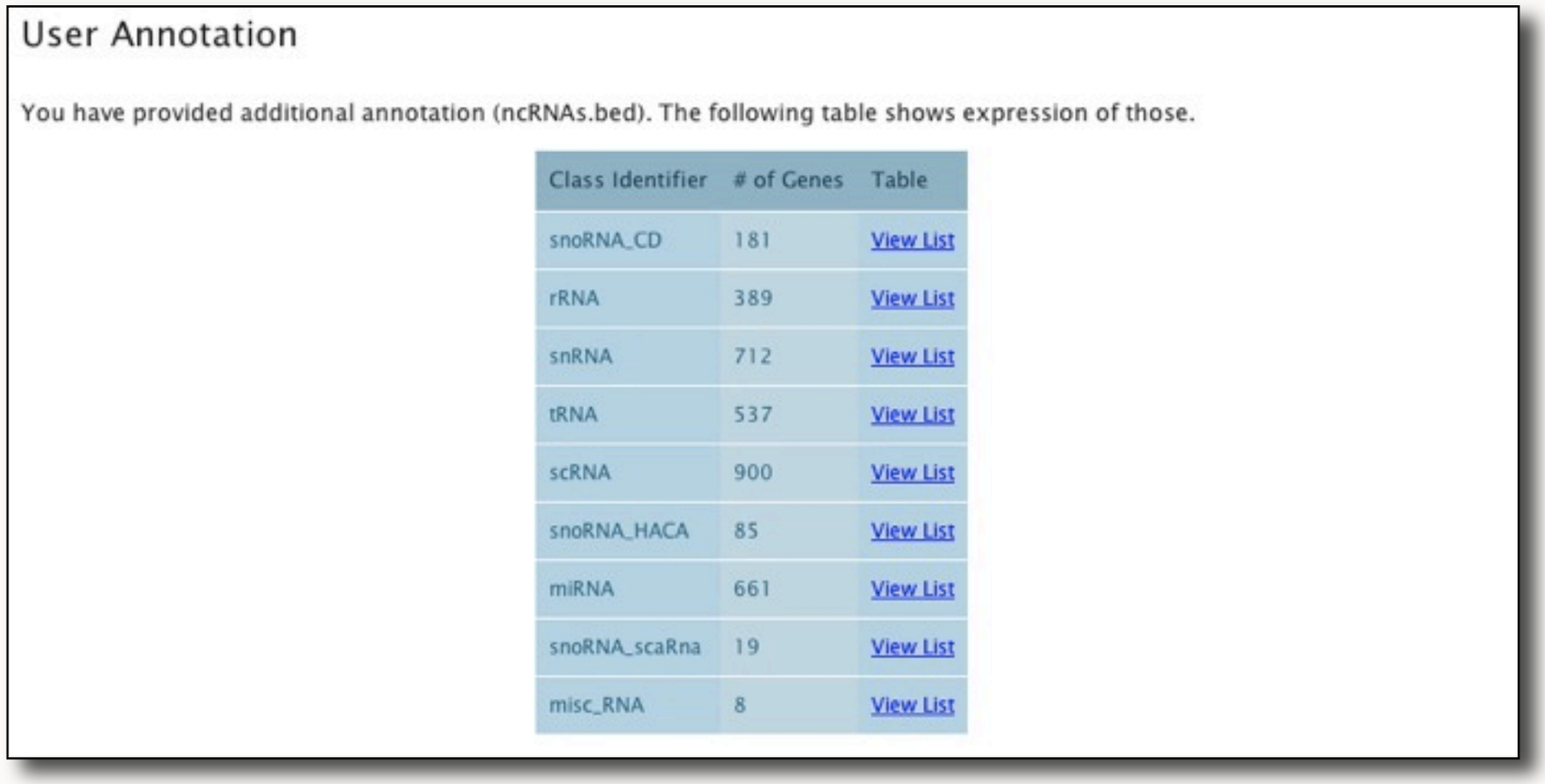


new candidate in intergenic region

new candidate antisense to an annotated miRNA



USER ANNOTATION



User Annotation

You have provided additional annotation (ncRNAs.bed). The following table shows expression of those.

Class Identifier	# of Genes	Table
snoRNA_CD	181	View List
rRNA	389	View List
snRNA	712	View List
tRNA	537	View List
scRNA	900	View List
snoRNA_HACA	85	View List
miRNA	661	View List
snoRNA_scaRna	19	View List
misc_RNA	8	View List

The output is in the same format as the analysis with the user defined annotations instead of the known ncRNA loci.

If you have not uploaded an annotation file, there will be no User Annotation section on the result page.

DOWNLOAD

Of course all the expression data and predicted candidates DARIO calculated are downloadable in bed format.

It is possible to use the predicted ncRNA candidates as ***User Annotation*** in upcoming DARIO runs.

We hope that you will enjoy working with DARIO!

Dario is a cooperative project of

David Langenberger
and
Mario Fasold